Approche statistique pour la définition du profil d'un utilisateur de système de recherche d'information

Lynda Tamine, Nesrine Zemirli, Wahiba Bahsoun

Approche statistique pour la définition du profil d'un utilisateur de système de recherche d'information

Lynda Tamine-Lechani, Nesrine Zemirli, Wahiba Bahsoun

Institut de Recherche en Informatique de Toulouse
Equipe Systèmes d'Information Généralisés
118 Route de Narbonne, 31062 Toulouse CEDEX 9
tamine,nzemirli,wbahsoun@irit.fr

Résumé

Un processus d'accès personnalisé à l'information a pour objectif de délivrer à l'utilisateur une information pertinente et appropriée à ses préférences, ses centres d'intérêts ou plus globalement son profil. Ce papier présente une méthode de définition du profil de l'utilisateur qui s'inscrit dans une approche statistique utilisant le comportement de l'utilisateur comme source permettant de définir implicitement son modèle. Cette méthode s'articule plus particulièrement sur l'interaction entre dimensions du profil représentées par l'historique des recherches et centres d'intérêt de l'utilisateur.

***Mots-clés** : accès personnalisé à l'information, profil utilisateur, contexte*

Abstract

The goal of a personnalized information access process is to carry out retrieval for each user incorporating his preferences, interests or more globally his profile. This paper presents a method for learning the user profile ; it's based on a statistical approach which exploits the user's behaviour in order to infer implicitly the corresponding model. The method focuses on the interaction of user's past search history and interests viewed as his profile dimensions.

***Key-words**: personalized information access, user profile, context*

1 INTRODUCTION

L'accès à une information pertinente, adaptée aux besoins et profil de l'utilisateur est un enjeu capital dans le contexte actuel caractérisé par une prolifération massive de ressources d'informations hétérogènes. Malgré les développements récents dans le domaine de la recherche d'information (RI), force

est de constater que les résultats produits par un moteur de recherche sont en deçà des attentes des 85% d'utilisateurs exploitant un moteur de recherche lors de leurs activités quotidiennes [38]. Les raisons évoquées portent essentiellement sur le volume et l'inintelligibilité des informations retournées [6, 29], leur désorientation et démotivation face à des interfaces qui induisent une surcharge cognitive. De récentes études [25] montrent que l'origine de ces problèmes réside dans le caractère non personnalisé du processus d'accès à l'information.
Dans ce cadre, la personnalisation est une dimension qui permet la mise en œuvre de systèmes centrés utilisateurs, non dans le sens d'un utilisateur générique mais d'un utilisateur spécifique et ce, en vue d'adapter son fonctionnement à son contexte précis. C'est pourquoi les travaux s'orientent actuellement vers l'adaptation du cycle de vie d'un processus d'accès à l'information à un utilisateur spécifique en vue de lui délivrer une information pertinente relativement à ses besoins précis, son contexte et ses préférences. C'est une direction de recherches, en l'occurrence la *RI contextuelle*, qui combine un ensemble de technologies et de connaissances portant sur la requête et le contexte de l'utilisateur, dans une même infrastructure et ce, dans le but de délivrer les réponses les mieux appropriées à son besoin en information [17]. Les applications sont diverses : systèmes de recommandation, systèmes de filtrage de messages et d'informations, systèmes d'apprentissage, systèmes de recherche d'information (SRI) etc.

Indépendamment de l'objectif applicatif visé, on identifie trois principaux aspects à promouvoir dans les systèmes d'accès personnalisé :
(a) capacité à identifier l'intention conceptuelle de l'utilisateur,
(b) flexibilité du processus de sélection de l'information en vue de s'adapter au contexte d'utilisation courant,
(c) intelligibilité des interactions utilisateur - système.
C'est précisément le premier aspect que l'on aborde dans le présent article. Notre objectif, dans ce cadre, est en effet de proposer une structure descriptive de l'utilisateur, appelée communément profil, qui permet de traduire deux principaux facteurs qui constituent des raisons fondamentales qui plaident pour la personnalisation de l'accès à l'information, à savoir que [40] :

- les utilisateurs ont des objectifs différents, des centres d'intérêts différents, en conséquence des perceptions différentes de la notion de pertinence,
- un même utilisateur peut avoir différents centres d'intérêts à différents instants.

A cet effet, on décrit l'utilisateur par un profil à deux dimensions [41]. La première représente l'historique de ses interactions avec le SRI. La seconde caractérise ses besoins récurrents en information ; elle est inférée et évolue à partir de la première dimension, et permet de définir différents centres d'intérêt. Ces derniers sont à intégrer, en perspective, dans une bibliothèque exploitée dans un processus d'appariement requête - document - utilisateur.

L'organisation retenue pour cet article est la suivante : la section 2 présente un aperçu de l'état de l'art portant sur les approches et techniques de définition du profil. Nous y mettons en évidence la position de notre contribution dans le domaine. La section 3 présente de manière détaillée notre approche pour la définition des dimensions descriptives du profil. Nous développons notamment les principes de construction et évolution de ces dimensions. La section 4 résume notre proposition et en présente les perspectives.

2 DÉFINITION DU PROFIL DE L'UTILISATEUR : APERÇU DES APPROCHES ET TECHNIQUES

La mise en œuvre naïve d'un modèle classique de recherche d'information suppose que l'utilisateur est complètement représenté par sa requête et que les résultats retournés pour une même requête sont identiques même si elle est exprimée par des utilisateurs différents. Les problèmes immédiats posés par une telle hypothèse sont notamment l'ambiguïté du sens des mots, l'impossibilité de sélectionner des sources opportunes et l'inintelligibilité des résultats [6]. En outre, ces problèmes sont d'autant plus accentués que les requêtes sont courtes (≈ 2.29 mots par requête) [39] et que les sources d'information sont volumineuses et hétérogènes. Les premières solutions apportées à ce type de problèmes et pouvant s'apparenter à la personnalisation, sont les techniques de reformulation de requêtes par injection de pertinence [32, 33]. Cependant, vu le contexte actuel lié au volume d'informations, ces techniques sont peu viables [18, 37]. Les travaux s'orientent actuellement vers une définition plus large de l'utilisateur permettant de l'exploiter dans la chaîne d'accès à l'information. Cette section présente les notions ayant émergé de cette direction de travaux puis en rapporte les principales contributions portant sur la définition du profil de l'utilisateur.

2.1 Notions de base

Les notions de *contexte* et *situation* sont en amont de l'interaction utilisateur-système d'accès à l'information. Ces notions ont été initialement introduites, sans distinction de sens, par les travaux de Saracevic [35] et Ingwersen [15]. Le contexte (ou situation) y est défini comme l'ensemble des facteurs cognitifs et sociaux ainsi que les buts et intentions de l'utilisateur au cours d'une session de recherche. Une tentative de distinction entre ces notions a fait l'objet d'autres travaux [1, 36, 9] qui précisent que le contexte couvre des aspects larges tels que l'environnement cognitif, social et professionnel dans lesquels s'inscrivent des situations liées à des facteurs tels que le lieu, temps et l'application au cours. C'est le sens générique du contexte qui a été largement exploré cette dernière décennie en RI contextuelle [21, 31, 16, 4]. Même si les auteurs ne convergent pas vers une même définition, on retrouve toutefois des dimensions descriptives communes telles que l'environnement

cognitif, le besoin mental en information, et l'interaction liée à la recherche d'information [10].
Par ailleurs, la notion de *profil* couvre, à notre connaissance ces mêmes aspects, à la seule différence que la définition initiale (réductrice de la définition actuelle du contexte) et l'usage de ce terme revient à la communauté en filtrage d'information.

En résumé, on peut définir globalement le contexte ou profil de l'utilisateur, dans le cadre d'une activité de recherche d'information, comme l'ensemble des dimensions qui permettent de décrire et/ou inférer ses intentions et perception de la pertinence. Les travaux en RI contextuelle abordent alors peu ou prou l'une ou l'autre de ces dimensions pour décrire l'utilisateur puis l'intégrer à terme dans les différentes phases du processus d'accès à l'information : reformulation de la requête, sélection des sources d'information et évaluation de la pertinence de l'information.

2.2 Processus de définition du profil de l'utilisateur

Le profil de l'utilisateur couvre des aspects larges tels que son environnement cognitif, social et professionnel qui déterminent ses intentions au cours d'une session de recherche [10]. La plupart des travaux actuels en RI contextuelle focalisent à juste titre, sur la représentation de l'aspect lié à ces intentions qualifiées de centres d'intérêts. Dans cette perspective, la modélisation du profil de l'utilisateur a pour objectif fondamental de représenter puis faire évoluer ses besoins en information à court et moyen terme. C'est une question qui pose la double difficulté de traduire les centres d'intérêt de l'utilisateur d'une part et faire émerger leur diversité d'autre part.
Le processus de définition du profil de l'utilisateur peut être caractérisé par trois phases. La première porte sur la représentation des unités d'information représentant le profil. La deuxième phase est liée à l'instanciation de ce modèle au cours d'une activité de recherche d'information. Enfin, la troisième phase concerne l'évolution du profil au cours du temps. Chacune de ces phases met en jeu des approches et techniques de représentation et/ou de construction résumées ci-après.

2.2.1 Représentation

Le modèle de base le plus communément utilisé pour la représentation des centres d'intérêt de l'utilisateur est le modèle vectoriel où chaque centre est représenté par une liste de termes représentatifs. Cependant, on distingue trois principales approches de représentation : ensembliste, sémantique et multidimensionnelle.

– *Représentation ensembliste :* le profil y est généralement formalisé comme des vecteurs de termes pondérés [6, 12] ou classes de vecteurs non hiérarchisées [25, 30] ou hiérarchisées [20] permettant de prendre en

compte des centres d'intérêt multiples. La représentation non hiérarchique considère les centres d'intérêt comme indépendants dans la description du profil ; leur éventuelle dépendance peut être prise en compte lors de l'intégration du profil dans la phase d'évaluation de pertinence des documents. En revanche, la représentation sous forme de hiérarchie de classes permet de traduire les relations de spécificité/généralisation entre les centres d'intérêt.

– *Représentation sémantique :* la représentation du profil met en évidence, dans ce cas, les relations sémantiques entre informations le contenant. La représentation est essentiellement basée sur l'utilisation d'ontologies [13, 7, 28, 24] ou des réseaux sémantiques probabilistes [22, 42].
Dans le cadre de cette approche, les centres d'intérêts de l'utilisateur sont appariés aux concepts des domaines de l'ontologie. Un profil est alors représenté en termes de concepts de l'ontologie intéressant l'utilisateur. Les ontologies de référence utilisées dans ce cadre sont basées sur la catégorisation en hiérarchie générale de Yahoo, Magellan, Lycos et ODP (Open Directory Project).

– *Représentation multidimensionnelle* : le profil est structuré selon un ensemble de dimensions, représentées selon divers formalismes [2, 5]. Les propositions de standards P3P [14] pour la sécurisation des profils ont défini des classes distinguant les attributs démographiques des utilisateurs (identité, données personnelles), les attributs professionnels (employeur, adresse, type) et les attributs de comportement (trace de navigation). Dans ce sens, on retrouve dans [2] un modèle de représentation du profil structuré en dimensions (ou catégories) prédéfinies : *catégorie de données personnelles, catégorie de données de la source, catégorie de données de livraison, catégorie de données de comportement* et *catégorie de données de sécurité*. L'auteur a proposé ce modèle dans le cadre du développement d'un service avancé de bibliothèque numérique (recherche et livraison personnalisée de l'information sur le Web) à l'aide du système *EUROgatherer*.

2.2.2 Construction

La construction du profil traduit un processus qui permet d'instancier sa représentation à partir de diverses sources d'information. Ce processus est généralement implicite et basé sur un procédé d'inférence du contexte et des préférences de l'utilisateur via son comportement lors de l'utilisation :

– d'un système d'accès à l'information [19, 13] : requêtes et documents jugés explicitement ou implicitement pertinents (consultés et/ou imprimés et/ou sauvegardés etc.)
– d'un navigateur web [25, 3] : liens explorés, dernières pages visitées etc.

– d'autres applications [6, 25, 12] : les application de bureautique, les outils de messagerie électronique etc.

Les informations extraites de ces sources sont organisées selon le modèle de représentation du profil à l'aide de différentes techniques. La plus répandue est celle basée sur l'analyse statistique du texte selon l'algorithme de Rocchio [32]. L'autre technique largement utilisée également dans les procédés de construction du profil, est celle de la classification appliquée aux informations collectées de l'utilisateur. On distingue plusieurs variantes dont la classification simple [25, 8], la classification hiérarchique [20] ou la classification Bayésienne [27].

2.2.3 Evolution

L'évolution des profils désigne leur adaptation à la variation des centres d'intérêt des utilisateurs qu'ils décrivent, et par conséquent, de leurs besoins en information au cours du temps. La phase d'évolution ne prend un sens que lorsque le profil a une structure pérenne, ce qui permet de distinguer les besoins à court terme, construits à partir de la session d'interaction courante, des besoins à long terme qui sont une réelle représentation des centres d'intérêt persistants de l'utilisateur.
A notre connaissance, peu de travaux ont exploré le problème de l'évolution du profil de l'utilisateur sous l'angle de la dimension temporelle (court terme, long terme). L'évolution est davantage abordée comme un problème de représentation de la diversité des domaines d'intérêt de l'utilisateur en utilisant des techniques de classification [30, 26, 25] ou heuristiques liées la notion de cycle de vie artificielle d'un centre d'intérêt [8] .

2.3 Position de l'approche proposée dans le domaine

Dans le cadre de notre approche, le profil de l'utilisateur comprend ses centres d'intérêt à court terme et long terme. Le profil repose sur une représentation à deux dimensions corrélées : historique des interactions et centres d'intérêts. Le processus de définition du profil est fondé sur l'interaction des phases de construction et d'évolution. Plus précisément, le profil est construit et évolue à partir des informations collectées sur les documents jugés implicitement ou explicitement pertinents lors des interactions de l'utilisateur avec un SRI. Nous utilisons pour cela des opérateurs d'agrégation d'informations ainsi qu'une méthode statistique qui permet de scruter le changement dans les centres d'intérêt de l'utilisateur, au cours du temps.
Comparativement aux travaux les plus proches cités ci-dessus, notre approche de définition du profil s'en distingue par les principaux points suivants :

– le profil comporte **deux dimensions formellement représentées selon la dimension temporelle**,

– le procédé de construction du profil considère l'**importance relative des informations collectées** compte tenu de l'historique des interactions de l'utilisateur,
– le procédé d'évolution du profil repose sur **l'interaction entre ses dimensions**, sans utilisation d'autres ressources telles que des ontologies ou des classifieurs de concepts. **Une méthode statistique** est déployée pour cela afin d'évaluer, au cours du temps, la corrélation entre les contextes associés à différentes sessions de recherche.

3 DÉFINITION DU PROFIL

Selon notre approche, le profil de l'utilisateur est multidimensionnel, décrit plus précisément, par deux dimensions. La première représente l'historique de ses interactions avec le SRI ; elle est exploitée pour inférer la seconde dimension représentée par les divers centres d'intérêt de l'utilisateur. Les deux dimensions évoluent corrélativement au cours du temps.
De manière sommaire, on définit le profil d'un utilisateur à l'instant s par $U = (H^s, I^s)$ où H^s représente l'historique des interactions de l'utilisateur jusqu'à l'instant s avec le SRI et I^s représente la bibliothèque de ses centres d'intérêt inférés jusqu'à l'instant s. Plus précisément, notre procédé de définition du profil se décline en un cycle comportant deux principales étapes.

La première étape consiste à représenter puis faire évoluer l'historique des interactions de l'utilisateur avec le SRI par agrégation des informations collectées à partir de ses sessions de recherche successives. Une session de recherche est particulièrement décrite par l'association d'une requête et d'un ensemble de documents associés, jugés explicitement ou implicitement par l'utilisateur.
La seconde étape a pour but de construire puis faire évoluer les centres d'intérêt de l'utilisateur en se basant sur la dimension *historique des interactions*. Plus précisément, on détermine des périodes d'apprentissage qui définissent des jalons pour l'extraction de centres d'intérêt à court terme, qualifiés de *contextes d'usage*, à partir des informations agrégées dans l'historique des interactions. L'évolution des centres d'intérêt est alors basée sur une mesure de corrélation des rangs qui évalue le degré de changement entre contextes d'usage associés à des périodes successives.
Les paragraphes qui suivent détaillent les principes de représentation, construction et évolution de l'historique des interactions et des centres d'intérêt de l'utilisateur.

3.1 Historique des interactions

Soit q^s la requête soumise par un utilisateur U à la session de recherche S^s se déroulant à l'instant s, et D^s l'ensemble des documents pertinents pour l'utilisateur durant cette session. Un document est considéré comme perti-

nent s'il a été ainsi jugé par l'utilisateur de manière explicite ou implicite. A cet effet, on se réfère à une catégorisation du comportement de l'utilisateur, traduisant des jugements implicites de pertinence, largement adoptée par les travaux du domaine [19]. Cette catégorisation construit une fonction *pertinence* qui associe à chaque action type (impression, lecture, sauvegarde etc.) un degré de pertinence.
Pour notre part, on utilise une fonction constante qui traduit davantage un indicateur de pertinence sûre. On note alors $R_u^s = \cup_{s_0..s} D^s$ l'ensemble des documents dèja *visités* et jugés pertinents par l'utilisateur lors des sessions de recherche passées depuis l'instant s_0. On propose dans ce qui suit l'utilisation de matrices pour la représentation d'une session de recherche et de l'historique des interactions.

3.1.1 Représentation d'une session de recherche

La session de recherche S^s est représentée par une matrice Document-Terme $D^s X T^s$ où T^s est l'ensemble des termes qui indexent les documents de D^s (T^s est une partie de l'ensemble des termes représentatifs des documents préalablement jugés pertinents noté $T\left(R_u^s\right)$)). Chaque ligne de la matrice S^s représente un document $d \in D^s$, chaque colonne représente un terme $t \in T^s$.

Dans le but d'affiner la représentation Document-Terme, on propose de décliner l'importance d'un terme relativement au profil de l'utilisateur dans le schéma de pondération terme-document. A cet effet, on calcule pour chaque terme t dans un document d à l'instant s, un coefficient de pertinence $CPT^s(t,d)$ qui traduit la pertinence relative d'un terme compte tenu des jugements de pertinence qu'il a émis et qui sont supposés êtres des indicateurs de son centre d'intérêt courant.
L'expression de ce coefficient est fondée sur l'hypothèse qu'un terme est d'autant plus important pour l'utilisateur qu'il cooccurre avec les termes qui lui sont *familiers* en ce sens qu'ils sont présents dans des documents dèja jugés. Les dépendances entre termes associés à des documents préalablement jugés sont vues comme des règles d'association [23].

Définition 1. *Le coefficient de pertinence d'un terme t dans un document d à l'instant s noté $CPT^s(t,d)$ est défini comme suit :*

$$CPT^s(t,d) = \frac{w_{td}}{l(d)} \cdot \sum_{t' \neq t, t' \in T(R_u^s)} cooc(t,t') \tag{1}$$

w_{td} est le poids du terme t dans le document d calculé selon le schéma classique tf*idf, $l(d)$ est la longueur du document d, $cooc(t,t')$ est le degré de confiance de la règle $(t \rightarrow t')$ quantifié à l'aide de la mesure EMIM (Expected Mutual Information Measure) [11], $cooc(t,t') = P(t,t') log \frac{P(t,t')}{P(t)P(t')}$, $P(t,t')$ est la proportion de documents contenus dans R_u^s indexés à la fois

par les termes t et t', $P(t)$ est la proportion de documents contenus dans R_u^s indexés par le terme t.
$S^s(d,t)$ est alors ainsi construit :

$$S^s = (CPT^s)^T \tag{2}$$

Où *T* est l'opérateur transposée de matrice

3.1.2 De la session de recherche à l'historique des interactions

L'historique des interactions de l'utilisateur est représenté par une matrice notée H^s de dimensions $|R_u^s| * |T(R_u^s)|$. Cette matrice est construite de manière incrémentale en ce sens qu'elle est mise à jour à chaque session de recherche en y reportant, par agrégation, les informations issues de la matrice S^s. A cet effet, on propose la définition d'un opérateur d'agrégation qui combine pour chaque terme son poids classique dans le document, calculé selon le schéma $tf * idf$, et ses poids atténués par les coefficients de pertinence calculés lors des sessions de recherche passées.

Définition 2. *L'opérateur d'agrégation des sessions de recherche, noté* $\oplus$*, est défini comme suit :*
$H^0(d,t) = S^0(d,t)$

$$H^{s+1}(d,t) = H^s(d,t) \oplus S^{s+1}(d,t) = \begin{cases} \alpha * w_{t,d} + \beta * S^{s+1}(d,t) \; si \\ t \notin T(R_u^{(s)}) \\ \alpha * H^s(d,t) + \beta * S^{s+1}(d,t) \; si \\ t \in T(R_u^{(s)}) et \; d \in R_u^{(s)} \\ H^s(d,t) \; sinon \end{cases} \tag{3}$$

$(\alpha + \beta = 1)$, $s > s_0$

La définition de l'opérateur $\oplus$ est fondée sur l'hypothèse que les termes associés aux centres d'intérêt de l'utilisateur sont récurrents. L'idée est alors d'affiner les descripteurs des documents dèja jugés par :

- expansion éventuelle avec des termes associés présents dans des documents pertinents,
- combinaison de l'importance classique de ses termes (relativement à la collection de documents) et de leur pertinence relative au profil, calculée à l'aide du coefficient $CPT(t,d)$ au cours des sessions de recherche passées.

3.2 Les centres d'intérêt

Les centres d'intérêt de l'utilisateur, contenus dans une bibliothèque notée I^s, constituent la seconde dimension de son profil traduisant ses besoins récurrents en information ; cette dimension est construite et évolue sur la base

de la première, en l'occurrence, l'historique de ses interactions.
La construction-évolution des centres d'intérêt est fondée sur une méthode cyclique qui procède en deux étapes. La première a pour objet d'extraire à partir de l'historique des interactions un centre d'intérêt candidat qualifié de *contexte d'usage*, qui traduit un besoin à court terme en information. L'objectif de la seconde étape est alors d'intégrer le contexte ainsi découvert dans la bibliothèque I^s en respectant l'hypothèse de diversité éventuelle des centres d'intérêt. Ceci traduit, à juste titre, la phase d'apprentissage des centres d'intérêt qui induit l'évolution du profil.

3.2.1 Extraction d'un contexte d'usage

A l'issue d'un cycle d'apprentissage représentant un nombre déterminé de sessions de recherche S^s agrégées dans l'historique H^s, est construit un contexte d'usage courant c^s défini comme suit :

Définition 3. *Un contexte d'usage traduit un besoin en information à court terme exprimé sur une courte période d'interactions avec le SRI ; il est représenté par un vecteur de termes pondérés, ordonnés par leur degré de représentativité du contexte ; pour chaque terme $t \in T(R_u^s)$, on calcule le poids associé comme suit :*

$$c^s(t) = \sum_{d \in R_u^s} H^s(d,t) \tag{4}$$

$c^s(t)$ est normalisé $c_n^s(t) = \frac{c^s(t)}{\sum_{t \in T(R_u^s)} c^s(t)}$.

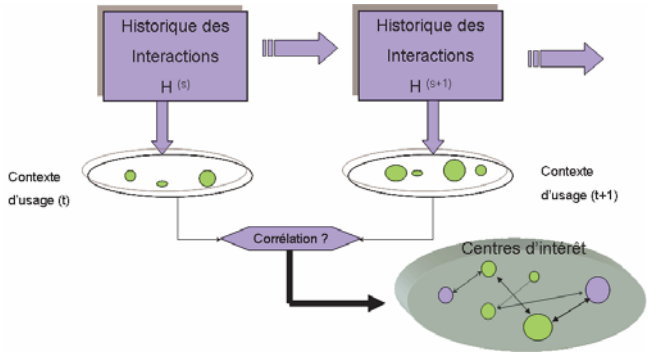
Un contexte d'usage est ainsi un vecteur extrait à partir de l'historique des interactions en sommant chaque colonne de la matrice associée.

3.2.2 Evolution des centres d'intérêt

On se base sur l'hypothèse qu'un utilisateur a divers centres d'intérêt et qu'il peut basculer d'un centre d'intérêt vers un autre au cours de sessions de recherche successives. Dans le but d'évaluer le degré de changement des contextes liés aux centres d'intérêt de l'utilisateur, on compare le contexte courant cc avec le contexte extrait à la période précédente pc comme illustré sur la figure 1.

Compte tenu de notre objectif, il est important de noter qu'on est davantage intéressés par l'évolution de l'ordre d'importance des termes dans la description des contextes d'usage extraits au cours des sessions, que par l'évolution de leur importance absolue. C'est pourquoi, nous adoptons une méthode statistique qui scrute le changement dans les centres d'intérêt à court terme, en utilisant le coefficient de corrélation des rangs de Kendall. Ce coefficient

FIG. 1 – Principe d'évolution des centres d'intérêt



permet en effet de savoir si deux variables aléatoires évoluent dans le même sens ou dans des sens différents autrement dits, non corrélés. On définit alors le degré de ce changement comme suit :

$$\Delta I = (cc \circ pc) = \frac{\sum_t \sum_{t'} S_{tt'}(pc) \cdot S_{tt'}(cc)}{\sqrt{(\sum_t \sum_{t'} S_{tt'}(pc)^2) \cdot (\sum_t \sum_{t'} S_{tt'}(cc)^2)}} \quad (5)$$

où $S_{tt'}(pc) = Signe(pc(t) - pc(t')) = \frac{pc(t)-pc(t')}{|pc(t)-pc(t')|}$, $S_{tt'}(cc) = Signe(cc(t) - cc(t'))$.

La valeur du coefficient ΔI est inscrite dans l'intervalle $[-1..1]$, où une valeur proche de -1 signifie que les contextes sont non similaires alors qu'une valeur proche de 1 signifie que les contextes sont proches sémantiquement. La distribution de ΔI est très correctement approchée par une loi de Laplace Gauss : $\Delta I \approx LG\left(0; \sqrt{\frac{2(2n+5)}{9n(n-1}}\right)$ dès que $n \geq 9$ [34] avec n, le nombre de termes traités, soit dans notre cas, $T(R_u^s)$. Ceci est un avantage pratique puisque le seuil critique de corrélation σ est alors donné par la table correspondante.

L'évolution des centres d'intérêt, qui a pour effet la mise à jour éventuelle de la bibliothèque I^s, est alors déterminé par le résultat de comparaison de ΔI relativement au seuil de corrélation σ. Plus précisément, la stratégie est la suivante :

1. $\Delta I > \sigma$. Les sessions de recherche sont inscrites dans le même contexte : pas d'indication sur l'évolution des centres d'intérêt de l'utilisateur ;
2. $\Delta I < \sigma$. Détection d'un changement de contexte ; deux configurations se présentent : découverte d'un nouveau centre d'intérêt ou évolution d'un autre préalablement découvert. On procède alors de la manière suivante :

– sélectionner $c^* = argmax^1{}_{c \in I^s}(c \circ cc)$,
– si $cc \circ c^* > \sigma$ alors :

  – affiner le descripteur du centre d'intérêt c^*,
  – mettre à jour la matrice H^s par élimination des lignes les moins récemment recalculées R^s_u,

– si $cc \circ c^* < \sigma$ alors :
  – élargir la librarie des centres d'intérêt ie $I^{s+1} = I^s \cup c^*$,
  – réinitialiser la matrice H^s de manière à privilégier l'apprentissage de ce nouveau centre d'intérêt, poser $s_0 = s$.

Les centres d'intérêt ainsi construits peuvent être réutilisés dans différentes étapes d'un processus d'accès personnalisé à l'information. Plus précisément, la bibliothèque I^s constitue alors une ressource pour :
– la réecriture de la requête,
– la mise en œuvre de l'appariement requête-document,
– l'ordonnancement des résultats de recherche

4 CONCLUSION

Dans la perspective d'un accès personnalisé à l'information, nous avons défini dans cet article, un profil multidimensionnel pour modéliser l'utilisateur. Nous introduisons la dimension temporelle afin de traduire tant l'évolution que la diversité des centres d'interêt de l'utilisateur au cours de ses interactions avec un SRI. Le profil est décrit au travers l'interaction de deux dimensions. La première traduit l'historique de ses interactions avec le SRI, représentée par une matrice issue de l'application d'un opérateur d'agrégation des informations collectées implicitement lors des sessions de recherche successives. La seconde dimension traduit les centres d'intérêt de l'utilisateur dérivés automatiquement à partir de l'historique des interactions. Le profil évolue selon une approche statistique basée d'une part, sur la distribution des termes dans les documents jugés explicitement ou implicitement pertinents et d'autre part, sur une mesure de corrélation permettant de scruter et traduire, au cours du temps, tant le changement que la diversité des centres d'intérêt de l'utilisateur.

L'approche de définition du profil ainsi présentée est essentiellement caractérisée par :

1. l'introduction de la dimension temporelle pour traduire l'évolution du profil,

[1]L'argument maximum ou Argmax représente la valeur de la variable pour laquelle la valeur de la fonction concernée atteint son maximum

2. la définition et utilisation d'une mesure de pertinence relative des termes pour un profil : cette mesure considère l'information véhiculée par les documents jugés par l'utilisateur,
3. l'exploitation de la seule dimension historique des recherches de l'utilisateur pour la construction et évolution de ses centres d'intérêt : aucune autre ressource n'est requise,
4. l'utilisation d'une méthode statistique pour maintenir la diversité des centres d'intérêt : cette méthode produit, à terme, une librarie qui peut être utilisée comme ressource pour personnaliser l'accès à l'information.

Des points méritent cependant d'être affinés. Le plus important porte sur la définition des périodes d'évolution des centres d'intérêt. En effet, la variation des centres d'intérêt de l'utilisateur, décelée à travers les requêtes qu'il a émises, ne présentent pas forcément des régularités prévisibles ; ainsi, la méthode statistique proposée serait confrontée à un risque d'erreur difficilement mesurable. Même si ce risque pourrait être amoindri en réduisant au mieux ces périodes, une perspective intéressante est de mener une réflexion plus poussée sur un compromis entre les différents paramètres qui régulent l'évolution des centres d'intérêt d'un utilisateur.

Par ailleurs, nous envisageons à court terme, de valider ces propositions dans le cadre d'une tâche d'évaluation inscrite dans les étapes du projet APMD (Accès Personnalisé à des Masses de Données). A cet effet, nous mettrons en oeuvre un cadre d'évaluation où seront définis des collections de test et mesures d'efficacité selon des protocoles proches de ceux définis dans le cadre de la campagne internationale d'évaluation en RI, en l'occurrence TREC .

5 REMERCIEMENTS

Cette recherche a été partiellement soutenue par le Ministère délégué à la Recherche et aux Nouvelles Technologies, dans le programme ACI Masses de Données, projet MD-33 (http ://apmd.prism.uvsq.fr/)

RÉFÉRENCES

[1] B. Allen. Information seeking in context : Proceedings of an international conference on research in needs, seeking and use in different context. pages 111–122, 1997.

[2] G. Amato et U. Staraccia. User profile modelling and applications to digital librairies. In *Proceedings of the 3rd European Conference on Research and avanced technology for digital libraries*, pages 184–187, 1999.

[3] R. Armstrong, D. Freitag et D. Joachims. A learning apprentice for the world wide web. AAAI Spring symposium on Information gathering from Heterogeneous, distributed environments, 1995.

[4] Bisson G. Bruandet M.F. Bottraud, J.C. Expansion de requêtes par apprentissage automatique dans un assistant pour la recherche d'information. In *Actes du congrès CORIA*, pages 89–105, Mars 2004.

[5] M. Bouzeghoub et D. Kostadinov. Personnalisation de l'information : Aperçu de l'état de l'art et définition d'un modèle flexible de définition de profils. In *Actes de la 2nde Conférence en Recherche d'Information et Applications CORIA*, pages 201–218, 2005.

[6] J. Budzik et K.J Hammond. User interactions with every applications as context for just-in-time information access. In *Proceedings of the 5th international conference on intelligent user interfaces*, pages 44–51, Mars 2000.

[7] V.K.R Challam. Contextual information retrieval using ontology based user profiles. In *Master of science in computer science*. Jawaharlal Nehru Technological University, 2004.

[8] K. Chen, L.and Sycara. Webmate : A personal agent for browsing and searching. In *Proceedings of the 2nd international conference on autonomous agents and multi agent systems, Minneapolis*, pages 10–13, 1998.

[9] C. Cool. The concept of situation in information science. *Annual review of information science and technology*, 35 :5–42, 2001.

[10] C. Cool et A. Spink. Issues of context in information retrieval : an introduction to the special issue. *In Journal of Information Processsing and Management (IPM)*, 38(55) :605–611, 2002.

[11] F. Crestani et E. van Risjbergen Cuttrel. A study of robability kinematics in information retrieval. In *ACM Transactions on information systems*, pages 225–255, 1998.

[12] S. Dumais, E. Cuttrel, J.J. Cadiz, G. Jancke, R. Sarin et D.C Robbins. Stuff i've seen : A system for a personal information retrieval and reuse. In *Proceedings of the 26th ACM SIGIR International Conference on Research and Development*, pages 72–79, 2003.

[13] S. Gauch, J. Chaffe et P. Pretschner. Ontology based user profiles for search and browsing. volume Special issue on user modelling for Web and hypermedia information retrieval, 2003.

[14] http ://www.W3C.com. 2005. 2005.

[15] P. Ingwersen. Cognitive perspectives of information interactions : elements of a cognitive information retrieval theory. *Journal of documentation*, 52(1) :3–50, 1996.

[16] P. Ingwersen et K. Jarvelin. Information retrieval in context. In *In Proceedings of the 27th ACM SIGIR Worshop on information retrieval in context*, pages 6–8, July 2004.

[17] Allan J et al. Challenges in information retrieval and langage modelling. In *Workshop held at the center for intelligent information retrieval*, Septembre 2002.

[18] B. Jansen, A. Spink et J. Bateman. Searchers : the subjects they search and sufficiency : a study of large sample of excite searches. In *Proceedings of the 1998 Web-net World conference of the WWW, Internet and Intranet*, pages 7–12. Faculty of science, University college Dublin, 1998.

[19] N. J Kelly. Understanding implicit feedback and document preference : a naturalistic study. In *PHD dissertation*. Ritgers University, New Jersey, January 2004.

[20] R. Kim et K. Chan. Learning implicit user interest hierarchy for context in personalization. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 101 – 108. ACM Press, 2003.

[21] S. Lawrence. Context in web search, ieee data engineering. *Bulletin*, 23(3) :25–32, 2000.

[22] C. Lin, G.R Xue, H.J Zeng et Y. YU. Using probabilistic latent semantic analysis for personalised web search. In *Proceedings of the APWeb Conference*, pages 707–711, 2005.

[23] S.H. Lin, C.S Shih, M.C Chen, J. Ho, M. Ko et Y. M. Huang. Extracting classification knowledge of internet documents with mining term-associations : A semantic approach. In *In the 21th International SIGIR Conference on Research end Development in Information Retrieval*, pages 241–249, 1998.

[24] F. Liu et C. Yu. Personalized web search for improving retrieval effectiveness. In *IEEE Transactions on knowledge Data Engineering*, volume 16, pages 28–40, 2004.

[25] J.P Mc Gowan. A multiple model approach to personnalised information access. In *Master thesis in computer science*. Faculty of science, University college Dublin, February 2003.

[26] S. Mizarro et C. Tasso. Ephemeral and persistent personalisation in adaptive information access to scholarly publications on the web. In *Proceedings of the 2nd International Conference on adaptive hypermedia and adptive Web-based systems*, pages 306–316, 2002.

[27] D. Mladenic. Personal webwatcher : design and implementation. In *Technical Report IJS-DP-7472*. J. Stefan Institute, Department for Intelligent Systems, 1998.

[28] N. Nanas, U. Uren et A. Deroeck. Building and applying a concept hierarchy representation of a user profile. In *Proceedings of the 26th Annual ACM Conference on Research and Development in Information Retrieval SIGIR*, pages 154–204, 2003.

[29] G. Nunberg. As google goes, so goes the nation. In *New York times*, May 1997.

[30] M. Pazanni, J. Muramatsu et D. Billsus. Syskill and webert : Identifying interesting web sites. In *Proceedings of the 13th National Conference on Artificial intelligence*, pages 54–61, 1996.

[31] J. Quiroga, L.M.and Mostafa. An experiment in building profiles in information filtering : the role of context of user relevance feedback. *Journal of Information Processing and Management (IPM)*, 38 :671–694, 2002.

[32] J. Rocchio. Relevance feedback in information retrieval. Englewood Cliffs, 1971. Prentice Hall.

[33] I. Ruthven et M. Lalmas. A survey on the use of relevance feedback for information access systems. volume 18, pages 95–145, 2003.

[34] G. Saporta. Probabilités, analyse de données et statistique. *Eds Technip*, 1990.

[35] T. Saracevic. The stratefied model of information retrieval interaction : extension and applications. In *Proceedings of the 60th annual meeting of the American Society for Information Science. Medford, NJ*, pages 313–327, 1997.

[36] D. H Sonnenwald. Evolving perspectives of human behaviors : contexts, situation, social networks and information horizons. In *Exploring the contexts of information behaviour : Proceedings of the 2nd international conference on reserach in information needs, seeking and use in different contexts*, pages 176–190, 1999.

[37] A. Spink, J. Bateman et B. Jansen. User's searching behaviour on the excite web search engine. In *Proceedings of the 1998 World Conference of the WWW, Internet and Intranet*, pages 7–12, 1998.

[38] A. Spink, B. Jansen, D. Wolfram et T. Saracevic. From e-sex to e-commerce : Web search changes. *IEEE Computer*, 35(3) :107–111, 2002.

[39] A. Spink, D. Wolfarm, M.B Jansen et T. Saracevic. Searching the web : the public and their queries. *Journal of American Science on Information and Technology (JASIST)*, 52(3) :226–234, 2001.

[40] J. Su et M. Lee. An exploration in personalized and context-sensitive search. In *Proceedings of the 7th annual UK special interest group for computatonal linguists research colloquium*, 2003.

[41] L. Tamine, M. Boughanem et N. Zemirli. Learning the user's interests using the search history. In *Proceedings of NIPS 2005 Workshop on Machine Learning for Implicit Feedback and User Modeling*, 2005.

[42] J.R Wen, N. Lao et W. Y Ma. Probabilistic model for contextual retrieval. In *Proceedings of the 27th annual internationsl ACM SIGIR Conference on Research and development in Information retrieval*, pages 57–63, August 2004.